349. Recent advances and future needs in genotype imputation

P.M. VanRaden*, D.J. Null and A.S. Al-Khudhair

USDA, Animal Genomics and Improvement Lab, Bldg 5 BARC-West, Beltsville, MD 20705, USA; paul.vanraden@usda.gov

Abstract

Genotype imputation has enabled selecting and using many more loci in research and prediction for much less cost than genotyping all individuals with the same array density or technology. Genotyping arrays are updated often to include new QTLs or higher-effect markers and applied to new animals without re-genotyping all previous or reference animals. That may require imputing the new loci from less-dense to higher-density arrays and from descendant genotypes to ancestors in the opposite direction of normal reference and target populations. Routine expansion and recycling of haplotype libraries and updating only individuals expected to change can speed imputation for routine predictions. This study compared imputation strategies for 78,964 loci of 4.6 million Holsteins. Sequence imputation is more challenging due to higher error rates and more complex variants, but new techniques could make sequencing cost-competitive with arrays for routine predictions.

Introduction

Genotypes at thousands of loci are combined into genomic relationships, and genomic predictions are usually obtained by multiple regression of phenotypes on genotypes. Solutions require replacing any unknown genotypes using allele frequencies, imputed genotypes, allele content (dosage), or genotype probabilities (Zheng *et al.*, 2011). Unknown phenotypes usually are easy to exclude from models but can also be 'imputed' to reduce computation using the same canonical transformation as when all traits are measured (Ducrocq and Besbes, 1993). Unknown genotypes can be imputed using pedigrees and a linear model one locus at a time from observed genotypes (Gengler *et al.*, 2007). Similar algebra uses pedigree relationships to linearly impute genomic relationships for ungenotyped animals from genotyped animals in single-step predictions (Aguilar *et al.*, 2010) and can improve marker effects by including phenotypes of ungenotyped parents, for example. More accurate haplotype-based imputation uses allele patterns across linked loci such as in long-range phasing or hidden Markov models and high-quality genome reference assemblies. This nonlinear imputation restricts genotype dosage to the range of homozygous reference to homozygous alternate (usually 0 to 2), whereas dosages from linear imputation can exceed the valid range. Known genotypes, pedigrees, and phenotypes could be used to impute the missing genotypes; however, accuracy is usually high by imputing genotypes just once without using the phenotypes, which often have low heritability and low correlations with individual genotypes. Accuracy is sometimes higher with multi-breed than single-breed imputation or using two steps instead of directly imputing all variants from the lowest density (VanRaden *et al.*, 2013). Imputation strategies for large datasets must balance accuracy with computing costs and adapt to properties such as array densities, sequence depth, error rates, and population structure in the input data. Imputation methods developed in animal breeding are often hundreds of times faster than from human genetics but with similar accuracy, especially with pedigree available (Sargolzaei *et al.*, 2014; Miar *et al.*, 2017). By processing in birth date order, most progeny haplotypes can be quickly selected from the two haplotypes of each genotyped parent or grandparents instead of from a long list of population haplotypes. This study provides a speed and accuracy comparison and a brief overview of genotype imputation strategies.

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_349, © P.M. VanRaden *et al.* 2022

Materials & methods

Microarray genotypes for 78,964 SNPs of 4.6 million Holsteins were imputed using Findhap version 3 either with no prior haplotype library or by obtaining priors from genotypes of a subset population that included 369,063 bulls plus their 73,813 dams, followed by one iteration including all animals. About 30,000 newly genotyped animals are imputed weekly using the prior haplotype library from the previous monthly update. The weekly imputation is nearly as accurate as the full monthly, with prediction correlations near 1.0 in larger breeds (Wiggans *et al.*, 2015). Each month the whole file is reprocessed to add about 120,000 new genotypes, update about 5,000 previous animals whose genotype or pedigree changed and obtain priors for the next weekly and monthly updates. Only the new and changing animals get their haplotype numbers updated, but any missing alleles within each haplotype can be filled by the added data and improve the imputed genotypes of other animals. The maximum length of haplotypes processed was reduced recently from 700 to 250 SNPs to speed imputation and limit the memory needed. The list of usable SNPs is updated about once per year. Routine evaluations with expanding reference populations must decide how frequently to reimpute all genotypes, such as when new QTLs are added or which animals to update, for example, only the most recent or those whose pedigrees change. Such strategies were compared using Holstein genotypes from the Council on Dairy Cattle Breeding (CDCB) database that included 48 different arrays ranging from <3,000 to >600,000 usable markers.

Results

Imputation using no prior haplotype library and four iterations would take 24 days with 30 processors and up to 500 Gb memory. Time was reduced to nine days by first obtaining a prior haplotype library from only the bulls' and dams' genotypes. Each week, about one hour is required to impute genotypes of new animals using the previous month's haplotype library. About 10 hours are required once per month to reprocess the whole file and include the new animals. Input genotypes averaged 72% unknown, whereas only 0.03% were still unknown after imputation, and 1.7% were half known with one allele still missing. Allele frequencies are later substituted for the missing alleles when estimating SNP effects. The list of 78,964 SNPs includes about 80 known QTLs already genotyped plus locations of six more discovered QTLs expected to be available on future arrays to allow imputing those without starting the imputation from scratch. Some recently selected high-effect SNPs from the high-density (HD) array are genotyped for only about 5,000 animals (0.1% of the population) currently. Still, imputation fills nearly all (>99%) of the missing alleles. Other SNPs from the HD array, especially on the X chromosome, were added to fill gaps when converting to the ARS-UCD1.2 cattle reference map (Rosen *et al.*, 2020). Such SNPs and QTLs are provided to genotyping laboratories and included when designing their future arrays.

Discussion

Advances in imputation allow combining various datasets, but highly accurate prediction may require identifying more of the QTLs and genotyping those in both the reference and candidates for selection instead of relying on markers. Array genotypes are typically so accurate that animal and plant breeders rarely store their quality scores based on distance from cluster centroids, for example. Uncertain genotypes are instead set to missing, and markers or samples with high missing or error rates are not used so that a low, uniform error rate of 1% or less may be assumed for all array genotypes remaining. That strategy allows efficient computation and storage of input data, but output files are much larger after imputation and include many uncertain genotypes. Sequence genotypes often are less accurate due to more read errors and both alleles not being observed across the whole genome, such as in regions with low coverage or where alignment to the reference assembly is challenging. Imputation to sequence can discover better markers or QTL, but high accuracy is needed for imputed sequence variants to outcompete highly linked array genotypes directly measured in the reference population. Rare or less frequent alleles may have poor imputation accuracy and may thus not improve prediction even if their actual effects are larger than

nearby genotyped markers (Zhang *et al.*, 2018). For example, the best linked marker from an array may be correlated by 0.97 with the true QTL but have nearly 100% correct genotypes compared to only 95% correctly imputed QTL genotypes for animals not sequenced. Reference cows may contribute much less to the reliability of prediction or marker selection if their sequence genotypes are imputed from only 6,000 to 20,000 usable SNPs and their individual phenotypes have low heritability. To obtain the gains in prediction accuracy expected from larger reference populations, well-designed genotyping strategies and accurate imputation are needed (Judge *et al.*, 2017).

Variant calling, phasing, and imputation can be combined so that each sample has only two haplotypes per region, and sequence read errors are suppressed instead of called as additional variants. Only two alleles are reported with array genotypes, whereas multiple alleles up to seven often are reported at the same locus in sequence data. Using such variants may require software redesign to exclude the less common alleles or track allele definitions within locations within chromosomes. A location can have all four bases (A, C, G, T) as SNPs plus multiple insertions and deletions of varying lengths. Copy number variants are harder to define, call, and impute than SNPs or indels and have not yet contributed much to genomic prediction (Chen *et al.*, 2021) despite some large, known effects. Standard variant call format (vcf) can require much space and be difficult to read: 'Do not write home-brewed VCF parsing scripts – it never ends well' (GATK Team, 2021). Simpler file formats may be possible that more languages could read with code of the user's choice. When merging separate vcf files, information about less common alleles can be lost for variants not reported if all samples in a single file are homozygous for the reference allele or if many samples have insufficient coverage to detect the allele. Instead of merging all raw sequence data, vcf files could each include and report genotypes and read depth at an agreed list of common locations. This strategy is like all genotyping arrays that include a standard set of 6,000 or 50,000 markers to simplify later merging with other arrays or files. The genomic vcf (gvcf) format instead takes more space to store approximate read depth for all genomic locations.

Low-coverage sequencing could soon become less expensive than arrays, allowing for genotyping many more variants with high accuracy in populations with good sequence reference panels available (Rubinacci *et al.*, 2021) and can also be used in developing those reference panels (VanRaden *et al.*, 2015). Methods to account for read errors in low-coverage data are needed (Ros-Freixedes *et al.*, 2020), and some of the read error bias can be overcome by aligning to both the reference genome and the alternate genome simultaneously (VanRaden *et al.*, 2019). The posterior genotype probabilities reported in vcf usually assume prior probabilities of 0.25, 0.5, and 0.25 and could be recalculated with actual genotype frequency priors (p^2, 2pq, and q^2) to obtain better dosage estimates. Distributed processing works well for arrays with accurate genotypes at predefined locations, but centralized processing may be needed for low-coverage sequences instead of sending imputed genotypes and updating those as reference populations improve and more breeds are added. Storage is a much bigger issue with raw sequence data and many rare alleles detected as more samples are sequenced. Genotypes could be stored for only the alternate alleles instead of storing all the homozygous reference genotypes. Similarly, referential compression can produce cram rather than bam files by storing only the differences of the raw reads from the reference assembly (Shi *et al.*, 2019). Larger populations have been genotyped for humans than livestock, and >300 million sequence variants have been imputed for >20 million humans (TOPMed Imputation Server, 2021). In animal genetics, the goal is not to routinely store or impute more variants but to identify the most useful variants that improve prediction at a reasonable cost. Advances in genotyping and imputation will continue to make genomic selection more affordable and accurate in future livestock populations.

References

Aguilar I., Misztal I., Johnson D.L., Tsuruta S., and Lawlor T.J. (2010) J. Dairy Sci. 93(2):743-752. https://doi.org/10.3168/jds.2009-2730

Chen L., Pryce J.E., Hayes B.J., and Daetwyler H.D. (2021) Animals 11:541. https://doi.org/10.3390/ani11020541

Ducrocq V and Besbes B (1993) J. Anim. Breed. Genet. 110, 81-92. https://doi.org/10.1111/j.1439-0388.1993.tb00719.x

GATK Team. (2021) VCF variant call format. https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format

Gengler, N., Mayeres P., Szydlowski M, *et al.* (2007) Animal 1:21-28. https://doi.org/10.1017/S1751731107392628

Judge M.M., Purfield D.C., Sleator R.D., and Berry D.P. (2017) J. Anim. Sci. 95(4):1489–1501. https://doi.org/10.2527/jas.2016.1212

Miar Y., Sargolzaei M., and Schenkel F.S. (2017) J. Dairy Sci. 100(4): 2837-2849. https://doi.org/10.3168/jds.2016-11590

Ros-Freixedes R., Battagin M., Johnsson M., Gorjanc G., Mileham A.J., *et al.* (2020) Genet. Sel. Evol. 50:64. https://doi.org/10.1186/s12711-020-00536-8

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., *et al.* (2020) GigaScience 9(3):giaa021. https://doi.org/10.1093/gigascience/giaa021

Rubinacci S., Ribeiro D.M., Hofmeister R.J., and Delaneau O. (2021) Nat Genet. 53:120–126. https://doi.org/10.1038/s41588-020-00756-0

Sargolzaei M., Chesnais J., and Schenkel F. (2014) BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478

Shi W., Chen J., Luo M., and Chen M. 2019. Bioinformatics 35(12):2058–2065. https://doi.org/10.1093/bioinformatics/bty934

TOPMed Imputation Server (2021) Nat. Inst. of Health https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home

VanRaden P.M., Bickhart D.M., and O'Connell J.R. (2019) J. Dairy Sci. 102(4):3216–3229. https://doi.org/10.3168/jds.2018-15172

VanRaden, P., Null, D., Sargolzaei, M., Wiggans, G., Tooker, M., Cole, J., *et al.* (2013). J. Dairy Sci. 96, 668–678. https://doi.org/10.3168/jds.2012-5702

VanRaden P.M., Sun C., and O'Connell J.R. (2015) BMC Genet. 16:82. https://doi.org/10.1186/s12863-015-0243-7

Wiggans G.R., VanRaden P.M., and Cooper T.A. (2015) J. Dairy Sci. 98(3):2039–2042. https://doi.org/10.3168/jds.2014-8868

Zhang Q., Sahana G., Su G., Guldbrandtsen B., Lund M.S., and Calus M.P.L. 2018. Genet. Sel. Evol. 50:62. https://doi.org/10.1186/s12711-018-0432-8

Zheng J., Li Y, Abecasis G.R., and Scheet P. (2011) Genet Epidemiol. 35(2):102–110. https://dx.doi.org/10.1002%2Fgepi.20552